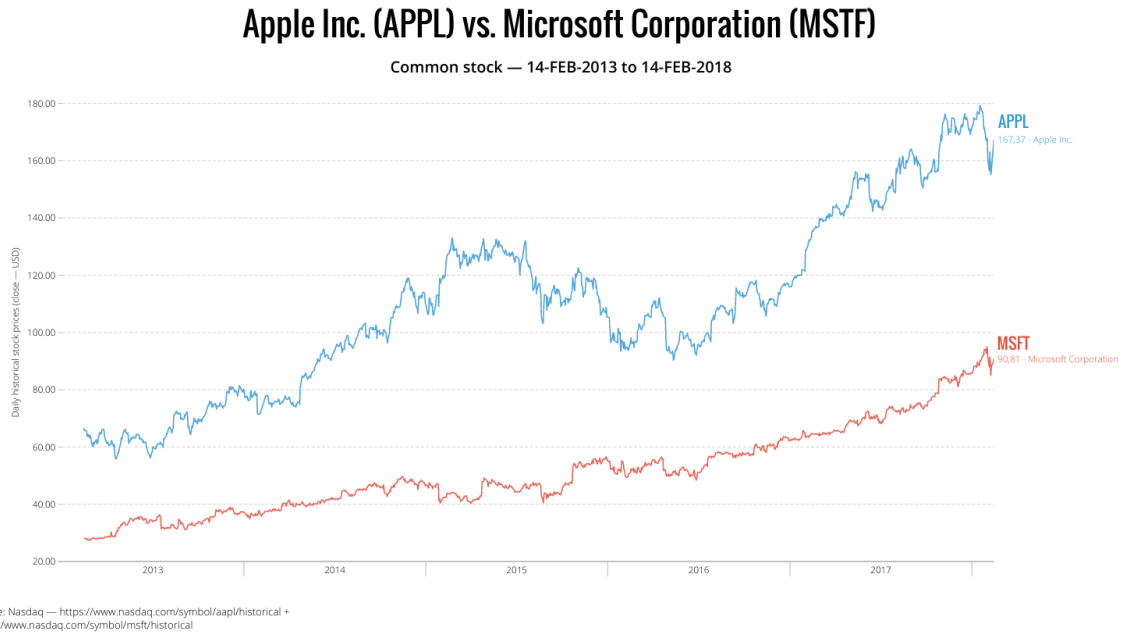
Rayyan Hassan

Interactive Data Visualization

Assignment 1

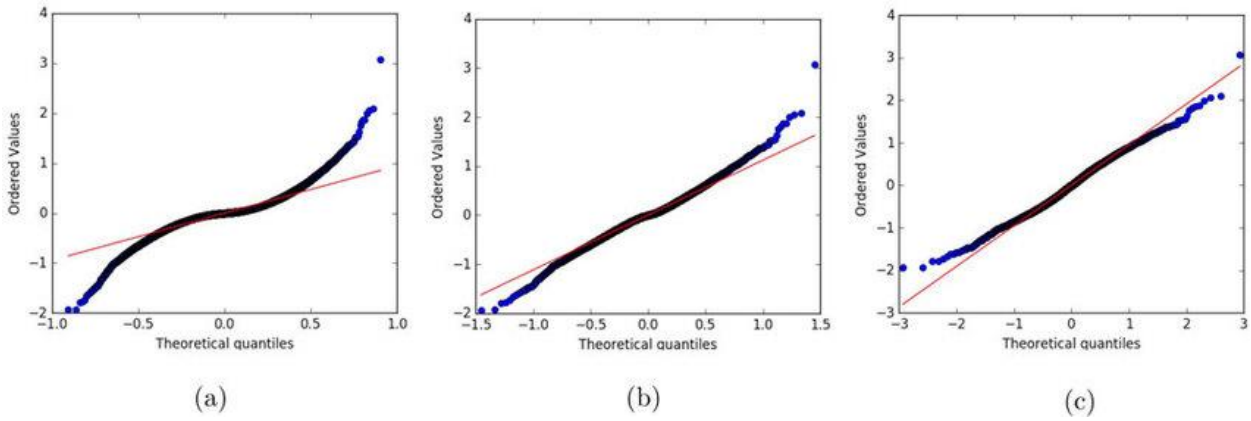
Data visualization has a myriad of advantages which make it an asset to use not only in data science but in most business ventures. These range from making the data easier to understand by presenting the data in a clear and cohesive way to recognizing patterns and trends which would otherwise have not been noticeable to the human eye.

Purpose of Visualization



We will first analyze, the Time-Series Data visualization which is an index chart of selected technology stocks, Apple and Microsoft from 2013–2018 in order to understand the main motivations behind this visualization and what it is accomplishing. Taking a deep look at the time series graph, it is possible to see the gain and loss that the stocks have accrued over a period of ten years. Its important to note that this visualization represent the prices of the various stock rather their relative gain and loss over that

period of time. This is quite important to note as usually investors are not concerned with the absolute price of the commodity but rather the change in price as in the percentage gain or loss that they will take if they invest in that commodity. Thus, they can perform the adequate risk management analysis before investing their money. Therefore, it might have been more beneficial for the analyst to transform the data in such a way as to get the general gain/loss rather than the absolute prices. Over a long enough period, it is also possible to see certain patterns and trends moreover, by keeping different technology stocks in the same graph, it is also possible to determine if there is some correlation between them.



(a) (b) (c)

Now for our second analysis I have chosen a Statistical Distribution which show the Q-Q plots of Mechanical Turk participation rates which is a Q-Q plot. When assessing data, it is helpful for an analyst to understand the statistical properties of the data thus they use the Q-Q plot to determine which kind of distribution does the data represent. They do this so that they can model their data to that particular model for a regression or classification task however they will obtain incorrect results if they choose the wrong model. The best representation of the data is given when the data is along the central diagonal line as that shows the best fit. Ideally, the data points should be along the straight line however that is only the case when the data points are entirely from one distribution which is rarely the case.

What information you have learned from the visualization?

From the index chart, I have gained a good idea about how technology has changed and evolved in the past decade as it is reflected in the price gain from 2013 to 2018. However, I believe that the time from 2014 was in actuality the most useful information present in the graph as it showed the existing

correlation between the different tech companies. While the relative gain and loss may have been different, it is easy to see that no negative correlation exists between the two companies.

From the Q-Q plot, I have learned that no one distribution can accurately represent the data points. It is quite probable that the data is not entirely uniform, gaussian, exponential etc but can be a combination of the aforementioned distributions. Usually in every data science problem, the distribution is unknown to the analyst however with this technique it is possible to gain a better understanding of the data that the analyst is modelling. By improving the understanding of the data, it is possible to use the relevant machine learning algorithms to perform the required task.

What information is hidden or not easy to tell from the visualization?

With respect to the time series visualization, there are quite a few hyperparameters that are hidden in the visualization. It does not take into consideration quite a lot of factors such as inflation, current events, development of technology, politics etc. and as such the patterns that emerge can be misleading without the appropriate knowledge. All time series visualizations are a combination of different features and hyperparameters and it is quite difficult to accurately determine what those features are simply by looking at the visualization. While it can be difficult to tell for novices in statistical analysis, there does exist a certain positive correlation between the different stocks as the rise and fall of the stocks happen during the same time period

While the Q-Q plot is quite useful when determining if two different data sets come from the same distribution, there are certain statistical qualities which are harder to see from the visualization alone. While in theory, it is possible to check shifts in location, shifts in scale, changes in symmetry, and the presence of outliers from this plot, it is either hidden or very difficult to see from the visualization that we were analyzing. It is however easy to see when two datasets have a different distribution from their distance from the vertical diagonal line. It is also difficult to verify whether they exhibit different tail behavior in this scenario.

Benefits and drawbacks of this type of visualization

When representing data, there are advantages and disadvantages of using the specific visualizations as no single visualization is perfect. The biggest advantage of using time series is data is its application for forecasting future predictions. It is very useful in regression problems in general and can predict them with a certain degree of accuracy. This is because through time series graphs it is possible to notice relevant patterns in the data such as there being rises in sales during Christmas period for example, Moreover, since the time series graphs require data over all intervals whether they be daily, weekly, monthly or even yearly, it becomes quite easy to clean the data as it is possible to notice the missing values. It can even show statistical information such as seasonality, autocorrelation, trends and stationarity.

While there are certain benefits to using this form of visualization, it does have its own share of limitations. One of the biggest limitations is the understanding of why the change is occurring in the data as the visualizations do not show the effect of the features or hyperparameters which are causing the change in the values. Thus, while we can see the effect in the visualization, we are unaware of the cause behind it. One of the biggest problems of using time series graphs is having the relevant data on hand and then extrapolating the missing values. In order to effectively use this form of visualization, it is key that the data be properly cleaned and transformed. Lastly, the observations are not mutually independent rather a single chance event can affect all later data points.

The Q-Q plots have quite a number of advantages when compared to other form of visualizations. It is possible to detect shifts in location, shifts in scale, changes in symmetry, and the presence of outliers from this plot. The sample sizes of the data set do not need to be equal for the plot to be implemented which is quite beneficial when the data points in the data are disproportionate. Furthermore, it can also verify if two data sets come from the same distribution and have similar tail behavior. Lastly, it allows us to determine which statistical distribution would be the ideal or close to ideal for the dataset which in turn allows us to better model our data.

There are certain drawbacks which come with using this plot. Unlike a histogram, it is very difficult to determine the number of peaks or its value in a Q-Q plot by simple human perception. Another limitation of this plot is that it needs a lot of data points as it becomes difficult to interpret the plot when the dataset is small.